Deep Learning Approximation Theory

Background & Universal Approximation

Ziyi Wang

Department of Statistics
Purdue University

February 17, 2021

Overview

1. Short Background
2. Elementary Approximation Theory
   - Stone-Weierstrass theorem
   - Main Result

Short Background

Definition

Shallow Network: A function $x \mapsto \sum_{j=1}^{m} a_j \sigma \left(w_j^\top x + b_j \right)$ is called basic shallow network.

Where *weight* matrix $W \in \mathbb{R}^{m \times d}$ and *bias* vector $v \in \mathbb{R}^m$ as $W_{j:} = w_j^\top$ and $v_j := b_j$

Definition

Basic deep network: Extending the matrix notation, given parameters
$w = (W_1, b_1, \ldots, W_L, b_L)$

$$f(x; w) := \sigma_L \left(W_L \sigma_{L-1} \left(\cdots W_2 \sigma_1 \left(W_1 x + b_1 \right) + b_2 \cdots \right) + b_L \right)$$

GOAL: Develop a model that can do well on the "Unseen" Data, thus need a measure of "Do well on the future Data"

Short Background

Definition (Empirical Risk)

Empirical Risk $\widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_i \ell\left(f\left(x_i\right), y_i\right)$.

Definition (Population Risk)

For future (random!) data, we consider (population) risk $\mathcal{R}(f) = \mathbb{E}\ell(f(x), y) = \int \ell(f(x), y)\mathrm{d}\mu(x, y)$

"Do well on future data" becomes "minimize $\mathcal{R}(f)$." This can be split into separate terms: given a training algorithm's choice $\hat{f}$ in some class of functions/predictors $\mathcal{F}$, as well as some reference solution $\bar{f} \in \mathcal{F}$,

$$
\begin{aligned}
\mathcal{R}(\hat{f}) = \mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f}) \quad & (\text{ generalization }) \\
+ \widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(\bar{f}) \quad & (\text{ optimization }) \\
+ \widehat{\mathcal{R}}(\bar{f}) - \mathcal{R}(\bar{f}) \quad & (\text{concentration/generalization}) \\
+ \mathcal{R}(\bar{f}); \quad & (\text{ approximation })
\end{aligned} \tag{1}
$$

Finite-width univariate approximation

Theorem

Suppose $g : \mathbb{R} \to \mathbb{R}$ *is* ρ *-Lipschitz. For any* $\epsilon > 0$*, there exists a 2-layer network* f *with* $\left\lceil \frac{\rho}{\epsilon} \right\rceil$ *threshold nodes* $z \mapsto 1[z \geq 0]$ *so that* $\sup_{x \in [0,1]} |f(x) - g(x)| \leq \epsilon$

Proof.

Define $m := \left\lceil \frac{\rho}{\epsilon} \right\rceil$ and $x_i := (i-1)\epsilon/\rho$, and Construct $\sum_i a_i 1\left[x - b_i\right]$ with

$$a_1 = g(0), a_i = g\left(x_i\right) - g\left(x_{i-1}\right), b_i := x_i$$

Then for any $x \in [0,1]$, pick the closest $x_i \leq x$, and note

$$|g(x) - f(x)| = \left|g(x) - f\left(x_i\right)\right| \leq \left|g(x) - g\left(x_i\right)\right| + \left|g\left(x_i\right) - f\left(x_i\right)\right| \tag{2}$$

$$= \rho(\epsilon/\rho) + \left| g\left(x_i\right) - g\left(x_0\right) - \sum_{j=2}^{i}\left(g\left(x_j\right) - g\left(x_{j-1}\right)\right) 1\left[x_i \geq x_j\right] \right| = \epsilon \tag{3}$$

Infinite width univariate approximation

Definition (infinite-width shallow network)

An infinite-width shallow network is characterized by a signed measure ν over weight vectors in $\mathbb{R}^p$:

$$x \mapsto \int \sigma\left(w^\top x\right) \mathrm{d}\nu(w)$$

Note: We can write any differentiable function as a form of infinite width network plus a constant.

$g : \mathbb{R} \to \mathbb{R}$ is differentiable, and $g(0) = 0$. If $x \in [0, 1]$, then $g(x) = \int_0^1 1[x \geq b] g'(b) \mathrm{d}b$

Multivariate approximation via bumps

Theorem

Let cont g and an $\epsilon > 0$ be given, and choose $\delta > 0$ so that $\|x - x'\|_\infty \leq \delta$ implies $|g(x) - g(x')| \leq \epsilon$. Then there exists a 3-layer network f with $\Omega\left(\frac{1}{\delta^d}\right)$ ReLU with $\int_{[0,1]^d} |f(x) - g(x)| \mathrm{d}x \leq 2\epsilon$

Note:

- Note the curse of dimension (exponential dependence on d).
- Proof involve two step approximation. First use step functions to approximate the target cont function, second use the two hidden layer network to approximate step funcions.

Lemma

Let g, δ, ϵ be given as in the theorem. For any partition $\mathcal{P}$ of $[0,1]^d$ into rectangles (products of intervals) $\mathcal{P} = (R_1, \ldots, R_N)$ with all side lengths not exceeding δ, there exist scalars $(\alpha_1, \ldots, \alpha_N)$ so that

$$\sup_{x \in [0,1]^d} |g(x) - h(x)|_{\mathrm{u}} \leq \epsilon$$

Multivariate approximation via bumps

Proof.

We will use the function $h = \sum_i \alpha_i 1_{Ri}$ from the lemma. Specifically, we will use the first two layers to approximate $x \mapsto 1_{Ri}(x)$ fol each i using $\mathcal{O}(d)$ nodes, and a final linear layer for the linear combination. Writing $\|f - g\|_1 = \int_{[0,1]} df(x) - g(x) \mid \mathrm{d}x$ for convenience, since

$$\|f - g\|_1 \leq \|f - h\|_1 + \|h - g\|_1 \leq \epsilon + \|h - g\|_1$$

and letting g_i denote the approximation to 1_{Ri},

$$\|h - g\|_1 = \left\| \sum_i \alpha_i \left(1_{R_i} - g_i\right) \right\|_1 \leq \sum_i |\alpha_i| \cdot \|1_{R_i} - g_i\|_1$$

so it suffices to make $\|1_{Ri} - g_i\|_1 \leq \frac{\epsilon}{\sum_i |\alpha i|}$. (If $\sum_i |\alpha_i| = 0$, we can set g to be the constant 0 network.) Let's do what we did in the univariate case, putting nodes where the function value changes.

Multivariate approximation via bumps

Proof.

For each $R_i := \times_{j=1}^{d} [a_j, b_j]$, and any $\gamma > 0$, define

$$g_{\gamma,j}(z) := \sigma\left(\frac{z-(a_j-\gamma)}{\gamma}\right) - \sigma\left(\frac{z-a_j}{\gamma}\right) - \sigma\left(\frac{z-b_j}{\gamma}\right) + \sigma\left(\frac{z-(b_j+\gamma)}{\gamma}\right)$$

and $g_\gamma(x) := \sigma\left(\sum_j g_{\gamma,j}(x_j) - (d-1)\right)$ (adding the additional ReLU layer is the key step!), whereby

$$g_\gamma(x) = \begin{cases} 1 & x \in R_i \\ 0 & x \notin \times_j [a_j-\gamma, b_j+\gamma] \\ [0,1] & \text{otherwise} \end{cases}$$

Since $g_\gamma \to 1_{R_i}$ pointwise, there exists γ with $\|g_\gamma - 1_{R_i}\|_1 \leq \frac{\epsilon}{\sum_i |\alpha_i|}$ □

Universal approximation with a single hidden layer

Definition (Universal Approximator)

A class of functions $\mathcal{F}$ is a universal approximator over a compact set S if for every continuous function g and target accuracy $\epsilon > 0$, there exists $f \in \mathcal{F}$ with

$$\sup_{x \in S} |f(x) - g(x)| \leq \epsilon$$

Notation Consider infinite-width networks with one hidden layer:

$$\mathcal{F}_{\sigma,d,m} := \mathcal{F}_{d,m} := \left\{ x \mapsto a^\top \sigma(Wx + b) : a \in \mathbb{R}^m, W \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m \right\}.$$

$$\mathcal{F}_{\sigma,d} := \mathcal{F}_d := \bigcup_{m \geq 0} \mathcal{F}_{\sigma,d,m}$$

Universal approximation with a single hidden layer

Theorem (Stone-Weierstrass theorem)

Let functions $\mathcal{F}$ be given as follows.

- *Each $f \in \mathcal{F}$ is continuous.*
- *For every x, there exists $f \in \mathcal{F}$ with $f(x) \neq 0$.*
- *For every $x \neq x'$ there exists $f \in \mathcal{F}$ with $f(x) \neq f(x')$ ($\mathcal{F}$ separates points).*
- *$\mathcal{F}$ is closed under multiplication and vector space operations ($\mathcal{F}$ is an algebra).*

Then for every continuous $g : \mathbb{R}^d \to \mathbb{R}$ and $\epsilon > 0$, there exists $f \in \mathcal{F}$ with $\|f - g\|_{\mathrm{u}} \leq \epsilon$.($\mathcal{F}$ is universal.)

Note Apply Stone-Weierstrass theorem(a very important theorem) will immediately give us that $\mathcal{F}_{cos,d}, \mathcal{F}_{ReLU,d}$ are universal approximators.

Proof of Stone-Weierstrass

Restate of the theorem

If $\mathcal{A}$ is a closed subalgebra of $C(X, \mathbb{R})$ that separates points, then either $\mathcal{A} = C(X, \mathbb{R})$ or $\mathcal{A} = \{f \in C(X, \mathbb{R})\} \mid f(x_0) = 0\}$ for some $x_0 \in X$

Lemma 1

Consider $\mathbb{R}^2$ as an algebra under coordinate addition and multiplication. The only subalgebras for $\mathbb{R}^2$ are $\mathbb{R}^2, \{(0,0)\}, \{(x,0) \mid x \in \mathbb{R}\}, \{(0,x) \mid x \in \mathbb{R}\}$, and $\{(x,x) \mid x \in \mathbb{R}\}$

To see that these are the only ones, consider a point $(a, b) \in \mathcal{A}$. If $\mathcal{A}$ contains a point such that $a \neq b \neq 0$, then (a, b) and $\left(a^2, b^2\right)$ are linearly independent. As a result, $\mathcal{A} = \mathbb{R}^2$. Now, the cases $a = b \neq 0$, $a \neq 0 = b$, or $a = 0 \neq b$ generate the other three nonzero subalgebras mentioned above. Finally, the only case remaining is if the only point happens when $a = b = 0$, which corresponds to the set $\{(0,0)\}$. Thus, the subalgebras mentioned above are the only possibilities.

Proof of Stone-Weierstrass

Lemma 2

For any $\epsilon > 0$ there is a polynomial P on $\mathbb{R}$ such that $P(0) = 0$ and $||X| - P| < \epsilon$ for $x \in (-1, 1)$

Proof. Let's start by considering the Maclaurin series for $f(t) = (1-t)^{\frac{1}{2}}$, given by $1 - \sum_{k=1}^{\infty} a_k t^k$, for constants a_k. Computing several derivatives, we see that

$$f^{(k+1)}(t) = \frac{(2k-1)f^{(k)}(t)}{2} \text{ for } k \geq 1. \text{ Therefore}$$

$$a_{k+1} = \frac{f^{(k+1)}(0)}{(k+1)!} = \frac{(2k-1)f^{(k)}(0)}{2(k+1)k!} = \frac{(2k-1)a_k}{2(k+1)}$$

Therefore, we have that

$$\lim_{k\to\infty} \left| \frac{a_{k+1}t^{k+1}}{a_k t^k} \right| = \lim_{k\to\infty} \frac{2k-1}{2k+2}|t| = |t|$$

Proof of Stone-Weierstrass

Thus, applying the ratio test, we see that the above series converges for $t \in (-1, 1)$. Now, let's show that this Maclaurin series actually equals $f(t)$. To see this, note that, according to Taylor's theorem, we know that the remainder for any Maclaurin polynomial of degree n must be given by $R_n(t) = \frac{f^{(n+1)(c)}}{(n+1)!} t^{k+1}$ for some $c \in (0, 1)$ Now, since $t \in (-1, 1)$ and $f^{(n+1)}(c)$ achieves it's maximum for $c = 0$, we see that this term must be less than a_{n+1}. Now, since the series above converges, we have that $\lim_{n\to\infty} R_n(t) = 0$, as required. To see that this series also converges for $t = 1$, we can apply the monotone convergence theorem to the counting measure on the natural numbers to conclude that

$$\sum_{k=1}^{\infty} a_k = \lim_{t\to 1} \sum_{k=1}^{\infty} a_k = 1 - \lim_{t\to 1} (1-t)^{\frac{1}{2}} = 1$$

Therefore, we have that Maclaurin Series for $f(t)$ converges to $f(t)$ for $t \in (-1, 1)$.

Proof of Stone-Weierstrass

This means that for every $\varepsilon > 0$ there exists a polynomial, $Q(t)$, such that $|f(t) - Q(t)| < \frac{1}{2}\varepsilon$. Substituting $t = 1 - x^2$, we see that

$$\left|f\left(1 - x^2\right) - Q\left(1 - x^2\right)\right| = ||x| - R(x)| < \frac{1}{2}\varepsilon$$

where $R(x)$ is the polynomial given by $Q\left(1 - x^2\right)$. Finally, let $P(x) = R(x) - R(0)$ Then, we have that

$$||x| - P(x)| < ||x| - R(x)| + |R(0)| < \varepsilon$$

where the last step follows from plugging $x = 0$ into the above inequality.

Proof of Stone-Weierstrass

Lemma 3

If $\mathcal{A}$ is a closed subalgebra of $C(X, \mathbb{R})$, then $|f| \in \mathcal{A}$ whenever $f \in \mathcal{A}$ and $\mathcal{A}$ is a lattice.

Proof. If $f = 0$, then $|f| = 0$, and therefore , $|f| \in \mathcal{A}$. Now, consider $f \neq 0$. Let $h : X \to [-1, 1]$ be given by $h = \frac{f}{\|f\|_u}$. Therefore, by lemma 2.9, for every $\epsilon > 0$ there exists a polynomial P such that $\||h| - P \circ h\|_u < \epsilon$. Since $h \in \mathcal{A}$ and P has no constant term, $P \circ h \in \mathcal{A}$. Now, since we have constructed a sequence whose limit if $|h|$ and $\mathcal{A}$ is closed, it follows that $|h| \in \mathcal{A}$. Thus, $|f| = \|f\|_u |h| \in \mathcal{A}$, as required. To see that $\mathcal{A}$ is a lattice, note that, by definition,

$$\max\{f, g\} = \frac{f+g+|f-g|}{2}$$
$$\min\{f, g\} = \frac{f+g-|f-g|}{2}$$

Therefore, by the first part of this lemma, we have that $\max\{f, g\}, \min\{f, g\} \in \mathcal{A}$

Proof of Stone-Weierstrass

Lemma 4

Suppose that $\mathcal{A}$ is a closed lattice in $C(X, \mathbb{R})$ and $f \in C(X, \mathbb{R})$. If for every $x, y \in X$ there exists $g_{xy} \in \mathcal{A}$ such that $g_{xy}(x) = f(x)$ and $g_{xy}(y) = f(y)$ then $f \in \mathcal{A}$

Proof. Let $\epsilon > 0$ be given. For all $x, y \in X$, define $U_{xy} = \{z \in X \mid f(z) < g_{xy}(z) + \epsilon\}$ and $V_{xy} = \{z \in X \mid f(z) > g_{xy}(z) - \epsilon\}$ and note that $x, y \in U_{xy}$ and $x, y \in V_{xy}$ Fix $y \in X$. Since, for all $x, x \in U_{xy}$, the set $\{U_{xy} \mid x \in X\}$ forms an open cover of X. Since X is compact, there exists a finite subcover, $\{U_{x_i y} \mid 1 \leq i \leq n\}$. Let $g_y = \max\{g_{x_1 y}, \ldots, g_{x_n y}\}$. Now, we have that $f < g_y + \epsilon$ over X and $f > g_y - \epsilon$ on $V_y = \cap_{i=1}^{n} V_{x_i y}$. Since, for all $y, y \in V_y$, the set $\{V_y \mid y \in X\}$ is an open cover for X Therefore, because X is compact, there exists a finite subcover, $\{V_{y_i} \mid 1 \leq i \leq k\}$. Let $g = \min\{g_{y_1}, \ldots, g_{y_k}\}$. From this we see that $\|f - g\|_u < \epsilon$. Since $\mathcal{A}$ is a lattice, it follows that $g \in \mathcal{A}$. Finally, since $\mathcal{A}$ is closed, we have that $f \in \mathcal{A}$

Proof of Stone-Weierstrass

Stone-Weierstrass Theorem.

Let $A_{xy} = \{(f(x), f(y)) \mid f \in \mathcal{A}\}$. Now, since $\mathcal{A}$ is a subalgebra of $C(X, \mathbb{R})$, A_{xy} is a subalgebra of $\mathbb{R}^2$. Therefore, by lemma 2.8 A_{xy} is either $\mathbb{R}^2, \{(0,0)\}, \{(x,0) \mid x \in \mathbb{R}\}, \{(0,x) \mid x \in \mathbb{R}\}$, or $\{(x,x) \mid x \in \mathbb{R}\}$. Now, since $\mathcal{A}$ separates points, A_{xy} cannot be $\{(0,0)\}$ or $\{(x,x) \mid x \in \mathbb{R}\}$. If $A_{xy} = \mathbb{R}^2$ then it follows from lemma 3 and lemma 4 that $\mathcal{A} = C(X, \mathbb{R})$. Finally, if A_{xy} is $\{(x,0) \mid x \in \mathbb{R}\}$ or $\{(0,x) \mid x \in \mathbb{R}\}$, then there exists some x_0 ($y = x_0$ or $x = x_0$ respectively) such that $f(x_0) = 0$ for all $f \in \mathcal{A}$. Furthermore, from lemma 4 and lemma 4, we have that $\mathcal{A} = \{f \in C(X, \mathbb{R})\} \mid f(x_0) = 0\}$. Finally, note that if $\mathcal{A}$ contains a constant function, then there does not exist an x_0 such that $f(x_0) = 0$ for all $f \in \mathcal{A}$. Thus, $\mathcal{A} = C(X, R)$

$\square$

Result of Stone-Weierstrass

Lemma 5

$\mathcal{F}_{\cos,d}$ is universal.

Proof. Let's check the Stone-Weierstrass conditions:

- Each $f \in \mathcal{F}_{\cos,d}$ is continuous.
- For each x, $\cos\left(0^\top x\right) = 1 \neq 0$.
- For each $x \neq x'$, $f(z) := \cos\left(\left(z - x'\right)^\top \left(x - x'\right) / \left\|x - x'\right\|^2\right) \in \mathcal{F}_d$ satisfies
$$f(x) = \cos(1) \neq \cos(0) = f\left(x'\right)$$
- $\mathcal{F}_{\cos,d}$ is closed under products and vector space operations as before.

Result of Stone-Weierstrass

Lemma 6

$\mathcal{F}_{\exp,d}$ is universal

Proof. Let's check the Stone-Weierstrass conditions:

- Each $f \in \mathcal{F}_{\exp,d}$ is continuous.
- For each x, $\exp\left(0^\top x\right) = 1 \neq 0$
- For each $x \neq x'$, $f(z) := \exp\left((z - x')^\top (x - x') / \|x - x'\|^2\right) \in \mathcal{F}_d$ satisfies

$$f(x) = \exp(1) \neq \exp(0) = f\left(x'\right)$$

- $\mathcal{F}_{\exp,d}$ is closed under VS ops by construction; for products,

$$\left(\sum_{i=1}^{n} r_i \exp\left(a_i^\top x\right)\right)\left(\sum_{j=1}^{m} s_j \exp\left(b_j^\top x\right)\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} r_i s_j \exp\left((a+b)^\top x\right)$$

Main Result

Theorem (Hornik, Stinchcombe, and White 1989.)

Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ *is continuous, and*

$$\lim_{z \to -\infty} \sigma(z) = 0, \quad \lim_{z \to +\infty} \sigma(z) = 1$$

Then $\mathcal{F}_{\sigma,d}$ *is universal.*

To prove this theorem we need two additional lemmas.

Lemma 7

Let F be a continuous squashing function and Ψ an arbitrary squashing function. For every $\varepsilon > 0$ there is an element H_ε of $\mathcal{F}_{\Psi,d}$ such that $\sup_{i \in R} |F(\lambda) - H_\varepsilon(\lambda)| < \varepsilon$

Main Result

Lemma 8

For every squashing function Ψ, every $\varepsilon > 0$, and every $M > 0$ there is a function $\cos_{M..} \in \mathcal{F}_{\Psi,d}$ such that

$$\sup_{i \in [-M+M|} |\cos_{M,x}(\lambda) - \cos(\lambda)| < \varepsilon$$

Proof Given $\epsilon > 0$ and continuous g, pick $h \in \mathcal{F}_{\cos,d}$ (or $\mathcal{F}_{\exp,d}$) with $\sup_{x\in[0,1]^d} |h(x) - g(x)| \leq \epsilon/2$. To finish, replace all appearances of cos with an element of $\mathcal{F}_{\sigma,1}$.

Note ReLU can be transformed as squashing function by

$$z \mapsto \sigma(z) - \sigma(z-1)$$

The End